THE IMPACT OF MISSING VALUE METHODS AND NORMALIZATION TECHNIQUES ON THE PERFORMANCE OF DATA MINING MODELS

MUNIRAH BINTI YAHYA

UNIVERSITI UTARA MALAYSIA
2011

THE IMPACT OF MISSING VALUE METHODS AND NORMALIZATION TECHNIQUES ON THE PERFORMANCE OF DATA MINING MODELS

A project submitted to Dean of Research and Postgraduate Studies Office in partial
Fulfillment of the requirement for the degree
Master of Science (Intelligent System)
Universiti Utara Malaysia

By
Munirah binti Yahya

KOLEJ SASTERA DAN SAINS
(College of Arts and Sciences)
Universiti Utara Malaysia

PERAKUAN KERJA KERTAS PROJEK
(Certificate of Project Paper)

Saya, yang bertandatangan, memperakukan bahawa
(I, the undersigned, certifies that)

MUNIRAH YAHYA
(806414)

calon untuk Ijazah
(candidate for the degree of) **MSc. (Intelligent System)**

telah mengemukakan kertas projek yang bertajuk
(has presented his/her project of the following title)

THE IMPACT OF MISSING VALUE METHODS AND NORMALIZATION TECHNIQUES ON THE PERFORMANCE OF DATA MINING MODELS

seperti yang tercatat di muka surat tajuk dan kulit kertas projek
(as it appears on the title page and front cover of project)

bahawa kertas projek tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan.
(that this project is in acceptable form and content, and that a satisfactory knowledge of the field is covered by the project).

Nama Penyelia
(Name of Supervisor) : **ASSOC. PROF. FADZILAH SIRAJ**

Tandatangan
(Signature) : _____ Tarikh (Date) : 28/2/2011

Nama Penilai
(Name of Evaluator) : **ASSOC. PROF. AZIZI ZAKARIA**

Tandatangan
(Signature) : _____ Tarikh (Date) : 28/2/2011

PERMISSION TO USE

In presenting this project in partial fulfillment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for copying of this project in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence by the Dean of Postgraduate and Research. It is understood that any copying or publication or use of this project or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my project.

Requests for permission to copy or to make other use of materials in this project, in whole or in part, should be addressed to

Dean of Research and Postgraduate Studies
College of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok
Kedah Darul Aman
Malaysia

ABSTRAK (BAHASA MALAYSIA)

Data yang banyak meyimpan pelbagai jenis rekod yang tidak diketahui dan ini menyukarkan proses analisa data. Nilai tersisih bagi data yang hilang atau data yang tidak lengkap boleh menyebabkan keputusan akhir yang diperolehi hasil daripada analisa tidak tepat. Lazimnya, masalah ini disebabkan tiada petunjuk tertentu bagi sesetengah analisis dalam menghasilkan keputusan akhir. Pembuatan keputusan yang berkauliti bergantung pada data yang berkualiti. Teknik prapemprosesan data menjadi dasar penting untuk menghasilkan data yang berkualiti terutamanya dalam perlombongan data. Jika perlombangan data juga tidak, ini akan menghasilkan keputusan perlombongan data yang tidak berkualiti. Lima langkah penting dalam prapemprosesan data ialah pembersihan data, penukaran data, pendiskretan data dan pengurangan data. Model untuk perlombongan data digunakan untuk menganalisis data secara meluas dalam penyelidikan. dan berupaya mengenal pasti perkaitan dan kesatuan dalam analisis. Regresi logistik ialah satu kaedah statistik yang penting bagi model perlombongan kerana kaedah ini berupaya meramal data yang dikategorikan. Kaedah lain yang digunakan untuk perlombongan data ialah *Neural Network* (NN) berjaya mengaplikasikan pembelajaran yang terselia dan pembelajaran yang tidak terselia. Kajian bertujuan untuk mengenal pasti teknik – teknik yang digunakan dalam pemprosesan data seperti memperbaiki data yang hilang dengan menggunakan dua teknik yang spesifik membaiki data yang hilang data iaitu *Mean of Attribute* dan *Mean of Each Target*. Hasil daripada eksperimen yang telah dijalankan, model Regresi Logistik menghasilkan keputusan iaitu ketepatan yang tinggi bagi proses pembelajaran model setelah menggunakan kaedah *Mean of Attribute*. Namun, eksperimen model NN bagi kedua-dua kaedah tersebut tidak memberi kesan ke atas model tersebut. Data yang digunakan untuk menjalankan proses pembelajaran perlu ditukar ke dalam bentuk yang boleh diterima sebagai input *Multilayer Perceptron* (MLP). Perbandingan beberapa teknik penormalan untuk setiap tiga set data iaitu teknik penormalan yang digunakan dalam kajian ini ialah kaedah penormalan Min-Max, penormalan Z-Score dan penormalan Sigmoidal. Untuk data Wisconsin Breast Cancer, penormalan Min-Max lebih sesuai. Bagi Pima Indians Diabetes dan Thyroid Disease, penormalan Sigmoidal lebih sesuai berbanding penormalan Min-max. Oleh itu hasil daripada eksperimen menunujkkan kaedah perlombongan data bukan hanya melibatkan kaedah membaiki data yang hilang dan kaedah penormalan, malahan jumlah data yang hilang turut member kesan keseluruhan set data.

Kata Kunci: Prapemprosesan Data, Data yang Hilang, Pernormalan, Model Perlombogan Data

ABSTRACT (ENGLISH)

In practice, the large datasets contain various types of anomalous records that significantly complicate the analysis problem. In particular, the prevalence of outliers, missing or incomplete data can completely invalidate the results obtained with standard analysis procedures, often with no indication that anything is wrong. High quality of decision making actually rely on high quality data, therefore data preprocessing has become the essential and important base of DM with no doubt because of no quality data, mean no quality mining results. Data preprocessing consists of interactive step such as data cleaning, data transformation, data reduction and data discretization. Data mining model have been used for extensive analysis in researches or data analysis work as it able to spot subtle relationships and associations. Logistic regression is an important statistical method for modeling and predicting categorical data. Another technique can be used in data mining task is neural network (NN) which have been successfully applied in a wide range of supervised and unsupervised learning applications. This study explored on the use of data preprocessing techniques such as missing values treatment namely *Mean of Attributes* and *Mean of Target*. The experimental results indicate that for the Logistic Regression models, models higher average accuracy is shown by data whose missing values were treated as *Mean of Attribute*. However, for NN models both missing value treatment did not affect the NN models. Prior to NNs training, the data needs to be transformed into form that is acceptable as input to Multi Layer Perceptron (MLP) network. Hence, several normalization techniques had been explored to compare which techniques suitable in each of the three datasets. There are several normalization techniques used for the experimental setup that is Min-Max normalization, Z-Score normalization and Sigmoidal normalization. For Wisconsin Breast Cancer data, Min-Max is preferable. However, for Pima Indians Diabetes and Thyroid Disease data set, Sigmoidal normalization is more preferable than the rest of the method. Hence, the experimental results indicate that the performance of DM models depends not only on the missing value and normalization techniques, it also depends on the amount of missing value in the whole data set.

Keywords: Data Preprocessing, Missing Values, Normalization, Data Mining Model

ACKNOWLEDGEMENT

By the Name of Allah, the Most Gracious and the Most Merciful

I would like to thank all people who have helped and inspired me during my study. I especially want to thank my supervisor Associate Professor Fadzilah Siraj, for her guidance during my research and study at Universiti Utara Malaysia. Her perpetual energy and enthusiasm in research had motivated all her advisees, including me. In addition, she was always accessible and willing to help her students with their research. As a result, research life became smooth and rewarding for me. I also want to thank to others lecturers who had taught me during the period of study. May all their valuable knowledge and information can help me to be a successful person.

My deepest gratitude goes to my family for their unflagging love and support throughout my life; this dissertation is simply impossible without them. I am indebted to my father, Yahya Bin Nadzam, for his care and love. He worked industriously to support the family and spare no effort to provide the best possible environment for me to grow up and attend school. He had never complained in spite of all the hardships in his life. I cannot ask for more from my mother, Badariah Binti Haji Rafaee, as she is simply perfect. Also for the other family members who always supporting and encouraging me to proceed till the end of my study. I would also like to express thanks to my beloved friends for their encouragement throughout the study. There is a tremendous sense of achievement in completing this study. It is certain that the project would not successful it did without their help and support.

Munirah Binti Yahya

College of Arts and Sciences

Universiti Utara Malaysia

2011

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In this chapter, the first section describes the context of the study that presents an introduction to data preprocessing stages in data mining (DM) approach using Logistic Regression models and Neural Network (NN) models, followed by the problem statement, the objectives of the study and the significance of the study. Finally, the scope of the study that includes the limitations of the study is also discussed.

1.1 DATA PREPROCESSING

Data are known as raw information in unorganized form which is limitless and present everywhere in the universe. These data are then stored in databases. To date, database growing rapidly, in fact very large collections continuously gathered about individuals, groups, government, companies and organizations which stockpiling very important personal data. The databases aim to help organizations manage human resources, better understand a market and customer behavior and although this information is collected will be used even to be shared to each other. Therefore, DM and knowledge discovery in database (KDD) are needed to deal with the flood of data.

In practice, the large datasets contain various types of anomalous records that significantly complicate the analysis problem. In particular, the prevalence of outliers, missing or

The contents of the thesis is for internal user only

BIBLIOGRAPHY

Abbass, H., A. (2002). *Artificial Intelligence in Medicine: An Evolutionary Artificial Neural Networks Approach for Breast Cancer Diagnosis.*pp.25.

Abonyi, J. & Feil, B. (2007). Clustering Techniques. In J. Abonyi & B. Feil (Eds.), *Cluster analysis for data mining and system identification* (pp. 8). Basel: Birkhkauser.

Agdas, D., & Ellis, R.D. (2010). Analysis of Temporary Traffic Control Cost Items in Transportation Construction Bidding Process. In Ruwanpura, J., Yaser Mohamed, Sang, H.L. (Eds.) *Proceedings of the 2010 Construction Research Congress. ASCE,* vol.2, pp.1103 – 1114.

Ali, K., & Warraich, M.A. (Eds.). (2010). A framework to implement data cleaning in enterprise data warehouse for robust data quality. Proceedings from ICIET'10: *International Conference on Information and Emerging Technologies.* 1 – 6. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Ali, M., Ghosh, M.K., & Al-Mamun, A. (Eds.). (2009). Multi-document Text Summarization: SimWithFirst Based Features and Sentence Co-selection Based Evaluation. Proceedings from ICFCC'09: *International Conference on Future Computer and Communication.* 93 – 96. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Athitsos, V., & Sclaroff, S. (2004). *Boosting Nearest Neighbor Classifiers for Multiclass Recognition.* Boston: University Computer Science Technology. (Report No, 2004-006).

Awad, E.M., & Ghaziri, H.M. (2007). Learning from Data. In E.M. Awad & H.M. Ghaziri (Eds.), Knowledge Management (pp. 338 – 340). Patparganj, Delhi: Pearson Education.

Balakrishnan, S., Narayanaswamy, R., Savarimuthu, N., & Samikannu, R. (Eds.). (2008). SVM ranking with backward search for feature selection in type II diabetes database. Proceedings from ICSMS'08: *IEEE International Conference on Systems, Man and Cybernetics.* 2628 – 2633. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Batmaz, I. & Koksal, G. (2010). Overview of Knowledge Discovery in Databases Process and Data Mining for Surveillance Technologies and EWS. In A. S. Koyuncugil & N.

Ozgulbas (Eds.), *Surveillance Technologies and Early Warning Systems: Data Mining Application for Risk Detection* (pp.1). United States: IGI Global.

Batyrshin, I.Z., & Sheremetova, L.B. (2007). Perception-Based Approach to Time Series Data Mining.

Bennett, K.P., Demiriz, A., & Maclin, R. (Eds.). (2002). Exploiting Unlabeled Data in Ensemble Methods. Proceedings from SIGKDD'02: *8th International Conference Knowledge Discovery and Data Mining at the Edmonton* (pp. 289 – 296). New York: ACM.

Berk, R. A. (2010). Data Mining within a Regression Framework. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 209). 2nd Edition. New York: Springer.

Bing, L. (2007). Web Data Mining. In L. Bing (Ed.), *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data* (pp.6). New York: Springer.

Brown, G. (2004). *Diversity in Neural Network Ensembles*. Birmingham: The University of Birmingham Press.

Buera, L., Miguel, A., Saz, O., Ortega, A., & Lleida, E. (Eds.). (2010). Unsupervised Data-Driven Feature Vector Normalization with Acoustic Model Adaptation for Robust Speech Recognition. Proceedings from TASL'10: *IEEE Transactions on Audio, Speech, and Language Processing*. 296 – 309. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Chai, D.J., Jin, L., Bae, K.H., Hwang, B., & Ryu, K.H. (Eds.). (2008). Continuous Sensor Data Mining Model and System Design. Proceedings from CIT'08: *IEEE 8th International Conference on Computer and Information Technology Workshops*. 501 – 506. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Chakrabarti, S., & Cox, E. (2008). Data Preprocessing. In S. Chakrabarti (Ed.) *Data mining: know it all*. United States: Morgan Kauffman. pp.78.
Dalamagkidis, K., Valavanis, K. P., & Piegl, L. A. (Eds.). (2010). Nonlinear Model Predictive Control with Neural Network Optimization for Autonomous Autorotation of Small Unmanned Helicopters. Proceedings from TCST'09: *IEEE Transactions on Control Systems Technology*. PP (99), 1 – 14. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Defit, S. (Ed.). (2009). Intelligent Data Transformation Based on Knowledge Based. Proceedings from ICIMT'2009: *International Conference on Information and Multimedia Technology* .393 – 395.

Dubitzky, W., Granzow, M., & Berrar, D.P. (2007). Data Mining. In W. Dubitzky, M. Granzow & D.P. Berrar (Eds.), *Fundamentals of Data Mining in Genomics and Proteomics* (pp.13). New York: Springer.

Dybowski, R. (2009). *Clinical Applications of Artificial Neural Networks*. Cambridge: Cambridge University Press.

Eggermont, J., Kok, J.N., & Kosters, W.A. (Eds.). (2004). Genetic Programming for Data Classification: Partitioning the Search Space. *The 19th Annual ACM Symposium on Applied Computing*. Cyprus: Hilton Park Hotel Nicosia. March 14 – 17, 2004.

Floares, A., Balacescu, O., Floares, C., Balacescu, L., Popa, T., & Vermesan, O. (Eds.). (2010). Mining knowledge and Data to Discover Intelligent Molecular Biomarkers: Prostate Cancer i-Biomarkers. Proceedings from SOFA'10: *4th International Workshop on Soft Computing Applications*. 113 – 118.

Floratou, A., Tata, S., & Patel, J.M. (Eds.). (2010). Efficient and Accurate Discovery of Patterns in Sequence Datasets. Proceedings from ICDE'10: *IEEE 26th International Conference on Data Engineering*. 461 – 472. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Frank, E., & Hall, M. (2003).Visualizing Class Probability Estimators. In Lavrac, N., Gamberger, D., Todorovski, L., Blockeel, H. (Eds.) *Proceedings of 7th European Conference on Principle and Practice of Knowledge Discovery in Databases (PKDD)*, Croatia, September 2003. Springer, pp. 168 – 179.

Gao, Q., Wang, X., Dong, C., & Li, C. (Eds.). (2010). Data Preprocessing for Prediction of Rerecirculating Water Chemistry Faults. Proceedings from ICCASM'10: *International Conference on Computer Application and System Modeling*. 14, 553 – 556. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Gao, X., & Wang, M. (Eds.). (2010). Applying data cleaning in Changqing Oilfield Company's Data Warehouse. Proceedings from IITA - GRS'10: *Second IITA International Conference on Geoscience and Remote Sensing*. 2, 605 – 607. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Georga, E., Protopappas, V., Guillen, A., Fico, G., Ardigo, D., Arredondo, M.T., *et.al.* (Eds.). (2009). Data mining for blood glucose prediction and knowledge discovery in diabetic patients: The METABO Diabetes Modeling and Management System. Proceedings from EMBC'09: *Annual International Conference of the IEEE on Engineering in Medicine and Biology Society*. 5633 – 5636. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Gopalan, N.P., & Sivaselvan, B. (2009). Data Transformation Techniques. In N.P. Gopalan & B. Sivaselvan (Eds.), *Data Mining and Techniques and Trends* (pp.13). New Delhi: Eastern Economy Edition.

Gunasundari, S., & Baskar, S. (Eds.). (2009). Application of Artificial Neural Network in identification of lung diseases. Proceedings from NaBIC'09: *World Congress on Nature & Biologically Inspired Computing*. 1441 – 1444. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Halkidi, M., & Vazirgiannis, M. (2010). Quality Assessment Approaches in Data Mining. In O. Maimon, L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 613).2nd Edition. New York: Springer.

Han, J. & Kamber, M. (2006). Data Cleaning. In J. Han & M. Kamber (Eds.), *Data mining: Concepts and Techniques* (pp.61). 2nd Edition. San Francisco: Morgan Kaufmann.

Janiszowski, K., Kobiecz, J., Patan, K., & Witczak, M. (2010). Process Modeling. In J. Kobiecz & J.M. Koscielny (Eds.), *Modeling, Diagnostics and Process Control: Implementation in the DiaSter System* (pp. 118). Berlin: Springer.

Jia, Y., Raja, A., & Ribarsky, W. (Eds.). (2010). Predictive Analytics Using a Blackboard-Based Reasoning Agent. Proceedings from WI-IAT'10: *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. 2, 97 – 100. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Jiansheng, W., Liangyong, H., & Xiaoming, P. (Eds.). (2010). A Novel Bayesian Additive Regression Trees Ensemble Model Based on Linear Regression and Nonlinear Regression for Torrential Rain Forecasting. Proceedings from CSO'10: *Third International Joint Conference on Computational Science and Optimization*. 2, 466 – 470. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Ji Liu; Musialski, Przemyslaw; Wonka, Peter; Jieping Y. (Eds.). (2009). Tensor completion for estimating missing values in visual data. Proceedings from ICCV'09: *IEEE 12th*

International Conference on Computer Vision. 2114 – 2121. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Kalatzis, F.G., Giannakeas, N., Exarchos, T.P., Lorenzelli, L., Adami, A., Decarli, M., et.al. (Eds.). (2009).Developing A Genomic-Based Point-of-Care Diagnostic System for Rheumatoid Arthritis and Multiple Sclerosis. Proceedings from IEMBS'09: *Annual International Conference of the IEEE on Engineering in Medicine and Biology Society*.827 – 830. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Khandelwal, V. (2009, March 31). *Neural Network*. Retrieved 16 December 2010, from http://www.articlesbase.com/electronics-articles/neural-network-842874.html#ixzz19TCe5L6D

Kholghi, M., Hassanzadeh, H., & Keyvanpour, M. (Eds.). (2010). Classification and Evaluation of Data Mining Techniques for Data Stream Requirements. Proceedings from 3CA'10: *International Symposium on Computer Communication Control and Automation*. 1, 474 – 478. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Kleinbaum, D.G., & Klein, M. (2010). Introduction to Logistic Regression. In D.G. Kleinbaum & M. Klein (Eds.), Logistic Regression: A Self-Learning Text (pp.129 – 302). 3rd Edition. New York: Springer.

Kretowski, M. & Czajkowski, M. (2010). An Evolutionary algorithm for Global Induction of Regression Trees. In Rutkowski, L., Sherer, R., Tadeusiewicz, R.(Eds.) Artificial Intelligence and Soft Computing, Part II: 10th International Conference (ICAISC), 2010. Springer, pp.157 – 164.

Lang, W., Morse, M., & Patel, J.M. (Eds.). (2009). Dictionary-Based Compression for Long Time-Series Similarity. Proceedings from TKDE'09: *IEEE Transactions on Knowledge and Data Engineering*. 22(11), 1609 – 1622. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Laurentis, G.D., Maino, R., & Molteni, L. (2010). Estimating a Model and Accessing its Discriminatory Power. In G.D. Laurentis, R. Maino & L. Molteni (Eds.) *Developing, Validating and Using Internal Ratings Methodologies and Case Studies* (pp. 213 – 334). United Kingdom: John Wiley & Sons Ltd.

Lean, Y., Shouyang, W., & Lai, K.K. (Eds.). (2006). An Integrated Data Preparation Scheme for Neural Network Data Analysis. Proceedings from TKDE'06: *IEEE Transactions on Knowledge and Data Engineering*. 18(2), 217 – 230. Retrieved 9

December 2010 from
http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Liang, Y., Rongzhan, L., Qian, S., & Ran, D. (Eds.). (2009). Application of Decision-Tree Algorithm to GSP Analysis and Assessment of Drug Safety Situation. Proceedings from the KAM '09: *Second International Symposium on Knowledge Acquisition and Modeling*. 2 , 103 – 106. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Li, L.W. (Ed.). (2009). Mining Regression-Classes in Fuzzy Point Data Sets. Proceedings from FSKD'09: *Sixth International Conference on Fuzzy Systems and Knowledge Discovery*. 2, 562 – 566. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Lisboa, P.J.G. (Ed.). (2002). A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural Networks* (Vol. 15, No.1, 11 – 39). Amsterdam: Elsevier Science Ltd.

Lisboa, P.J.G. (Ed.). (2004). A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural Networks* (Vol. 15, No.1, 11 – 39). Amsterdam: Elsevier Science Ltd.

Maio, D.A., Kay, S.M., Farina, A. (Eds.). (2010). On the Invariance, Coincidence, and Statistical Equivalence of the GLRT, Rao Test, and Wald Test. Proceedings from TSP'09: *IEEE Transactions on Signal Processing*. 58(4),1967 – 1979. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Middel, A. (2009). Estimating Residential Building Types from Demographic Information at a Neighborhood Scale. In G. Steinebach & Subhrajit Guhathakurta (Eds.), *Visualizing Sustainable Planning* (pp.195). New York: Springer.

Minjie, B., Jiaoxiong, X., & Jun, X. (Eds.). (2010). Database Preprocessing with AHP. Proceedings from FSKD'10: *Seventh International Conference on Fuzzy Systems and Knowledge Discovery*. 6, 2805 – 2809. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Moghaddam, B., & Shakhnarovich, G. (2003). Boosted Dyadic Kernel Discriminants. In Becker, S., Thrun, S., Obermayer, K. (Eds.) *Proceedings of the 2002 Conferences on Advances in Neural Information Processing System 15*. United States, Massachusetts Institutes of Technology, pp.761 – 767.

Nan, J., & Zhiqiang, C. (Eds.). (2010). Post mining of non-redundant association rules for sensor data estimation. Proceedings from: ICETC'10: *2nd International Conference on Education Technology and Computer*. 5, 102 – 106. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Nirkhi, S. (Ed.). (2010). Potential use of Artificial Neural Network in Data Mining. Proceedings from ICCAE'10: *The 2nd International Conference on Computer and Automation Engineering*. 2, 339 – 343. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Norsarini Salim. (2004). Discusses Using Neural Networks to Aid Diagnosis of Several Common Medical Conditions. *Medical Diagnosis Using Neural Networks.*

Pasolli, L., Melgani, F., & Blanzieri, E. (Eds.). (2010). Gaussian Process Regression for Estimating Chlorophyll Concentration in Subsurface Waters From Remote Sensing Data. Proceedings from LGRS'09: *IEEE Transaction on Geoscience and Remote Sensing Letters*. 7(3), 464 – 468. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Pattichis, C.S., & Pattichis, M.S. (Eds.). (2001). Adaptive Neural Network Imaging in Medical Systems. Proceedings from ACSSC'01: *Conference Record of the Thirty-Fifth Asilomar Conference on Signals, Systems and Computers*.1, 313 – 317. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Priddy, K.L., & Keller, P.E. (2005).Data Normalization. In K.L. Priddy & P.E. Keller

(Eds.), *Artificial Neural Networks: An Introduction* (pp.16).United States: SPIE.

Qing, A., Weidong, W., Boya, Z., Jing, L., & Kaiyuan, L. (Eds.). (2010). Application of Data Mining Based on Clinical Medicine Database. Proceedings from ICSPS'10: *2nd International Conference on Signal Processing Systems*. 3, 719 – 723. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Qing, A., Weidong, W., Zhiwen, L., & Kaiyuan, L. (Eds.). (2010). Explored research on data Preprocessing and Mining Technology for clinical Data Applications. Proceedings from ICIME'10: *The 2nd IEEE International Conference on Information Management and Engineering*. 327 – 330. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Quiroga, O., Melendez, J., & Herraiz, S. (Eds.). (2010). Sequence Pattern Discovery of Events Caused by Ground Fault Trips in Power Distribution Systems. Proceedings from the ICHQP'10:
14th International Conference on Harmonics and Quality of Power. 1 – 6. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Raymer, M.L., Doom, T.E., Kuhn, L.A., & Punch, W.F. (Eds.). (2003). Knowledge Discovery in Medical and Biological Datasets Using a Hybrid Bayes Classifier/Evolutionary Algorithm. *IEEE Transactions on Systems, Man, and Cybernetics*. Part B, 33. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Rongguo, Z., Baihua, X., & Chunheng, W. (Eds.). (2010). Data Transformation of the Histogram Feature in Object Detection. Proceedings from ICPR'10*: 20th International Conference on Pattern Recognition*. 2893 – 2896. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Rub, G. & Brenning, A. (2010). Spatial Variable Importance Assessment for Yield Prediction in Precision Agriculture. In Cohen, P. R., Adams, N.M., Berthold, M.R. (Eds.) *Proceedings of 9th International Symposium (IDA), 2010 in Advances in Intelligent Data Analysis IX*, Tucson, AZ, USA, May 2010. Springer, pp.184.

Ruibin, Xi., Nan, L., & Yixin, C. (Eds.). (2009). Compression and Aggregation for Logistic Regression Analysis in Data Cubes. Proceedings from TKDE'08: *IEEE Transactions on Knowledge and Data Engineering*. 21(4), 479 – 492. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Sattler, K.U., & Schallehn, E. (Eds.). (2001). A Data Preparation Framework Based on a Multidatabase Language. Proceedings from the IDEAS'01: *International Symposium on Database Engineering & Applications*. 219 – 228. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Seetha, M., Sunitha, K.V.N., Lalitha Parameswari, D.V., & Ravi, G. (Eds.). (2010). Accuracy assessment of object oriented and knowledge base image classification using P-trees. Proceedings from ICCAE'10: *The 2nd International Conference on Computer and Automation Engineering*. 5, 760 – 763. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Seifert, J. W. (2004). *Data Mining: An Overview*. Washington: Congressional Research Service, Library of Congress.

Shamim, A., Shaikh, M.U., & Malik, S.U.R. (Eds.). (2010). Intelligent Data Mining in Autonomous Heterogeneous Distributed Bio Databases. Proceedings from ICCEA'10: *Second International Conference on Computer Engineering and Applica*tions. 1, 6 – 10. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Shahriar, M.S., & Anam, S. (Eds.). (2008).Quality Data for Data Mining and Data Mining for Quality Data: A Constraint Based Approach in XML. Proceedings from FGCNS'08: *Second International Conference on Future Generation Communication and Networking Symposia.*
2, 46 – 49. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Shmueli, G., Patel, N.R., & Bruce, P.C. (2010). Evaluating Classification and Predictive Performance. In G. Shmueli, N.R. Patel & P.C. Bruce (Eds.), *Data Mining for Business Intelligence: Concepts, Techniques, and Application in Microsoft Office Excel with XLMiner* (pp. 96). 2nd Edition, New Jersey: John & Wiley & Sons, Inc.

Srinath Srinivasa. (2010). *Data Mining, Data Warehousing and Knowledge Discovery Basic Algorithms and Concepts.* URL: http://www.scribd.com/doc/38999575/Data-Mining-Data-Warehousing-and-Knowledge-Discovery Accessed date: 18 February 2011.

Sung H. H., & Zhen, Y. Z. (2010). Clinical Informatics to Diagnose Cardiac Diseases Based on Data Mining. In Khuri, S., Lhotska, S., Pisanti, N. (Eds.) *Proceedings of First International Conference in Information Technology in Bio- and Medical Informatics (ITBAM), 2010*, Bilbao, Spain September 2010. Springer, pp.67- 77.

Sung, H.H., & Zhenyu, Z. (Eds.). (2010). Collaborative Intelligence for Intelligent Diagnosis Systems in Hospital Environment. Proceedings from WKDD '10: *Third International Conference on Knowledge Discovery and Data Mining.* 532 – 535. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Sykacek, P., & Roberts, S.J. (2003). Adaptive Classification by Variational Kalman Filtering. In Becker, S., Thrun, S., Obermayer, K. (Eds.) *Proceedings of the 2002 Conferences on Advances in Neural Information Processing System 15.* Massachusetts Institutes of Technology, pp.753 – 760.

Tang, K., Ponnuthurai Suganthan, N., Xi, Y., & Kai, Q.A. (Eds.). (2005). Linear Dimensionality Reduction Using Relevance Weighted LDA. *Pattern Recognition.* 38, 485 – 493.

Tarling, R. (2009).Multinomial Logistic Regression. In R. Tarling (Ed.) *Statistical Modeling for Social Researchers Principle and Practice* (pp.79). Oxon: Taylor & Francis e-Library.

Taylor, B.J. (2006). Automated Test Generation for Testing Neural Networks Systems. In B.J. Taylor (Ed.), *Methods, and Procedures for the Verification and Validation of Artificial Neural Networks* (pp.229 – 230). New York: Springer.

Tripathy, A., Das, S., & Patra, P.K. (Eds.). (2010). An improved design Approach in Spatial Databases Using frequent Association Rule Mining Algorithm. Proceedings from IADCC'10:
IEEE 2nd International Advance Computing Conference. 404 – 409. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Vijayabhanu, R., & Radha, V. (Eds.). (2010). Recognition and Elimination of Missing Values and Outliers From An Anaerobic Wastewater Treatment System Using K-Means Cluster. Proceedings from the ICACTE'10: *3rd International Conference on Advanced Computer Theory and Engineering*. 4, 186 – 190. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Vo, N., Vo, D., Challa, S., & Moran, B. (Eds.). (2009). Parametric Subspace Analysis for Dimensionality Reduction And Classification. Proceedings from the CIDM '09: *IEEE Symposium on Computational Intelligence and Data Mining*. 363 – 366. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Wang,B., Zhang, M., Zhang, B., & Wei, W. (Eds.). (2006). Data Mining Application to Syndrome Differentiation in Traditional Chinese Medicine. Proceedings from the PDCAT '06: *Seventh International Conference on Parallel and Distributed Computing, Applications and Technologies*.128 – 131. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Wasilewska, A. (Ed.). (2010). Descriptive Granularity. Proceedings from the GrC'10: *IEEE International Conference on Granular Computing*. 37 – 38. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Xiaohua, H. (2003). *DB-HReduction: A Data Preprocessing Algorithm for Data Mining Applications*. Retrieved 17 February 2011 from http://www.ischool.drexel.edu/faculty/thu/research-papers/aml3730.galley.pdf

Xiaoyong, C., Li, D., Qiang, Y., & Ling, C.X. (Eds.). (2004). Test-Cost Sensitive Naive Bayes Classification. Proceedings from the ICDM'04: *Fourth IEEE International*

Conference on Data Mining. 51 – 58. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Xinlin, Q., Liping, D., Deren, L., Pingxiang, L., Shi, L., & Liefei, C. (Eds.). (2009). Data Cleaning Approaches In Web2.0 VGI Application. Proceedings from GEOINFORMATICS'09: *17th International Conference on Geoinformatics*. 1 – 4. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Xiuyan, P., Zhonghui,H., & Tengfei, L. (Eds.). (2010). Neural Network Based Combining Prediction Model and Its Application in Ship Motion Prediction. Proceedings from WCICA'10: *8th World Congress on Intelligent Control and Automation*. 5624 – 5627. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Xu, J.L., Wei, S., &Meng, C.Z. (Eds.). (2010). Likelihood Function-Based Modulation Classification in Bandwidth-Constrained Sensor Networks. Proceedings from ICNSC'10: *International Conference on Networking, Sensing and Control*. 530 – 533. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Yan, H., & Diao, X. (Eds.). (2008). The Design and Implementation of Data Cleaning Knowledge Modeling. Proceedings from the KAM'08: *International Symposium on Knowledge Acquisition and Modeling*.177 – 179. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Yinghui, H., & Guanyu, L. (Eds.). (2010). Descriptive Models for Internet Of Things. Proceedings from ICICIP'10: *International Conference on Intelligent Control and Information Processing*. 483 – 486. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Young, T., Fehskens, M., Pujara, P., Burger, M., & Edwards, G.(Eds.). (2010). Utilizing Data Mining To Influence Maintenance Actions. Proceedings from AUTEST'10: *IEEE Transaction on AUTOTESTCON*. 1 – 5. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Yu, L., Pan, Y., & Wu, Y. (Eds.). (2009). Research on Data Normalization Methods in Multi-Attribute Evaluation. Proceedings from the CISE'09: *International Conference on Computational Intelligence and Software Engineering*. 1 – 5. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Yun, Y., & Ke, C. (Eds.). (2011). Temporal Data Clustering via Weighted Clustering Ensemble with Different Representations. Proceedings from TKDE'10: *IEEE Transactions on Knowledge and Data Engineering*. 23(2), 307 – 320. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Zhang, P., & Li, C. (Eds.). (2009). Automatic Text Summarization Based On Sentences Clustering and Extraction. Proceedings from the ICCSIT'09: *2nd IEEE International Conference on Computer Science and Information Technology*. 167 – 170. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Zhang, Y., Li, W., & Chen, Y. (Eds.). (2009). The Study of Multidimensional-Data Flow of Fishbone Applied for Data Mining. Proceedings from the SERA'09: *7th ACIS International Conference on Software Engineering Research, Management and Applications*. 86 – 91. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Zhen, W., & Zhou, M.X. (Eds.). (2008). Evaluating the Use of Data Transformation for Information Visualization. Proceedings from the TVCG'08: *IEEE Transactions on Visualization and Computer Graphics*. 14(6), 1309 – 1316. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Zhi, H.Z., & Yuan,J. (Eds.). (2004). NeC4.5: neural ensemble based C4.5. Proceedings from the TKDE'04: *IEEE Transactions on Knowledge and Data Engineering*. 16 (6), 770 – 773. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Zhu, S., Hong, W., Wu, Q., & Li, M. (Eds.). (2009). Research On Data Mining Model Of GIS-Based Urban Underground Pipeline Network. Proceedings from the ICCA'09: *IEEE International Conference on Control and Automation*. 1515 – 1520. Retrieved 9 December 2010 from http://ieeexplore.ieee.org.eserv.uum.edu.my/Xplore/dynhome.jsp

Zulaiha Othman, Azurazliza Abu Bakar, Zalinda Othman, & Suzanna Rosli. (2009). Development of the Preprocessing Agent's Knowledge for Data Mining Using Rough Set Theory. In Peng, W., Yuefeng, L., Polkowski, L., Yiyu, Y., Tsumoto, S., Guoyin, W. (Eds.) *Proceedings of 4th International Conference in Rough Sets and Knowledge Technology (RSKT), 2009*, Gold Coast, Australia, July 2009.Springer, pp.86 – 93.